

Inside the clustering window for random linear equations

Pu Gao*

School of Mathematical Sciences
Monash University
jane.gao@monash.edu

Michael Molloy†

Department of Computer Science
University of Toronto
molloy@cs.toronto.edu

Abstract

We study a random system of cn linear equations over n variables in $\text{GF}(2)$, where each equation contains exactly r variables; this is equivalent to r -XORSAT. Previous work has established a clustering threshold, c_r^* for this model: if $c = c_r^* - \epsilon$ for any constant $\epsilon > 0$ then with high probability all solutions form a well-connected cluster; whereas if $c = c_r^* + \epsilon$, then with high probability the solutions partition into well-connected, well-separated *clusters* (with probability tending to 1 as $n \rightarrow \infty$). This is part of a general clustering phenomenon which is hypothesized to arise in most of the commonly studied models of random constraint satisfaction problems, via sophisticated but mostly non-rigorous techniques from statistical physics. We extend that study to the range $c = c_r^* + o(1)$, and prove that the connectivity parameters of the r -XORSAT clusters undergo a smooth transition around the clustering threshold.

1 Introduction

The study of random constraint satisfaction problems (CSP's) has been revolutionized by a collection of hypotheses, arising from statistical physics, concerning the geometry of solutions. According to these hypotheses, before reaching the satisfiability threshold of a random CSP (e.g. r -SAT), the geometry of its solution space undergoes several phase transitions. Roughly speaking: there are specific constants, including the *clustering threshold*, the *freezing threshold*, and the *condensation threshold*, all below the satisfiability threshold (see [35] for an overview). These phase transitions indicate dramatic changes in the degree of the correlation between solutions as the density of a random CSP grows; which shed insights into why it is challenging to determine the satisfiability threshold for many CSP's, or to find efficient CSP solvers when the density is close to the satisfiability threshold. This paper focuses on the clustering threshold. When the density (the ratio of the number of constraints to n , the number of variables) of a random CSP instance exceeds the clustering

*Major part of this research was done when the author was affiliated with University of Toronto, supported by an NSERC Postdoctoral Fellowship.

†Research supported by an NSERC Discovery Grant and Accelerator Supplement.

threshold, the set of solutions a.a.s.¹ partitions into exponentially many clusters, whereas before the density reaches the clustering threshold, all solutions form a single cluster. One can move throughout any cluster by making small local changes; i.e. changing the values of $o(n)$ variables in each step. But solutions in two different clusters must differ globally; i.e. they differ on a linear number of variables.

While these hypotheses are, for the most part, not rigorously proven, they come from some substantial mathematical analysis. They explain many phenomena, most notably why some random CSP's are algorithmically very challenging (eg. [1, 22]). Intuition gained from these hypotheses has led to some very impressive heuristics (eg. Survey Propagation[37, 39, 4]), the best of the random r -SAT algorithms whose performance has been rigorously proven [11]), and some remarkably tight rigorous bounds on various satisfiability thresholds [13, 14, 15, 17]. Ding, Sly and Sun recently used an approach outlined by these hypotheses to prove the k -SAT conjecture [23], with a determination of the k -SAT satisfiability threshold for all large k . It is clear that, in order to approach many of the outstanding challenges around random CSP's, we need to understand clustering.

Amongst the commonly studied random CSP's, r -XORSAT (a.k.a. linear equations over $\text{GF}(2)$) is the one for which the clustering picture is, rigorously, the most well-established. The exact satisfiability threshold was established for $r = 3$ in [21], and then for $r \geq 4$ in [20, 45]. The clustering threshold c_r^* , and the structure of the solution clusters were analyzed in [10, 38] and then established rigorously in [27, 2]. These papers provide a very thorough description of the clusters for any constant density $c \neq c_r^*$. However, the birth of the r -XORSAT clusters is not well understood. There has been no description of the solution space when the number of constraints is $c_r^*n + o(n)$. This is the main target of this paper.

Consider a random process where random r -XORSAT constraints are added one at a time; this corresponds to a random hypergraph process where random r -uniform hyperedges are added one after the other. We show that the geometric structure of the whole solution space, specifically a key connectivity parameter, transits rather smoothly around clustering. Analysing the manner of a transition near the threshold of a phase transition is a common goal; see e.g. the extensive work on the birth of the giant component[5, 33, 34, 28], and the 2-SAT transition[6].

The cluster structure of r -XORSAT is simpler than that of most other models, and this has enabled researchers to prove challenging results for it long before proving them for other models. However, the structure of the clusters in the other models are hypothesized to be a generalization of the simpler structure in r -XORSAT. Understanding cluster properties in r -XORSAT often helps to predict properties in other CSP's. For instance, many CSP's have the generic property that, after the freezing threshold, most clusters consist of a solution σ to a subset (of linear size) of the variables, called *frozen variables*, along with all extensions of σ to the rest of the variables [3, 41, 4, 47]. This is true of r -XORSAT clusters, although in a simpler way: the set of frozen variables is invariant among clusters, whereas in other models, they can differ. Insights from the work on r -XORSAT have been valuable when studying more complicated models; eg. ideas from [2] led to [41, 42].

¹ A property holds a.a.s. (asymptotically almost surely) if it holds with probability tending to 1 as $n \rightarrow \infty$.

2 Main results

Before stating the main results, we formalise the concepts of clustering and connectivity of clusters mentioned in Section 1. We also give formal definitions of probability spaces under the discussion of this paper.

2.1 Random linear equations and random hypergraphs

Our model for a random system of equations is $X_r(n, m)$ defined as follows. We have n variables over $\text{GF}(2)$ and m equations each formed in the following way: The LHS is the summation of a uniformly chosen r -tuple of variables and the RHS is chosen uniformly from $\{0, 1\}$. We focus on the case $m = cn$ where $c = \Theta(1)$. We restrict ourselves to the case $r \geq 3$, as the case $r = 2$ behaves very differently, and is already well-understood (see eg.[6]). For this range of r, m , a simple first moment bound shows that a.a.s. no two equations will have the same r -tuple of variables, so choosing the r -tuples with or without replacement has a negligible effect; to be specific, we choose them without replacement.

It is not hard to see that this is equivalent to choosing an instance of r -XORSAT on n variables by uniformly choosing m r -tuples to form clauses, and then signing the variables within each clause uniformly at random. An assignment to the variables is satisfying if every clause contains an odd number of true literals.

A common alternate model is to choose each of the $\binom{n}{r}$ r -tuples of variables independently with probability p , and then form an equation for each r -tuple (with a uniformly random RHS). By conditioning on the “typical” number of chosen r -tuples, our results for $X_r(n, m)$ immediately translates to this model.

Given a system of linear equations over $\text{GF}(2)$, its *underlying hypergraph* is defined as follows: the vertices are the variables and for each equation, the set of variables appearing in that equation form a hyperedge. So the underlying hypergraph of $X_r(n, m)$ is distributed as $\mathcal{H}_r(n, m)$, the random r -uniform hypergraph on n vertices and m hyperedges uniformly chosen without replacement from the $\binom{n}{r}$ hyperedges in the complete hypergraph.

2.2 Clustering

We formalise the concept of solution clustering in CSPs. Given a CSP instance F , let $\Phi(F)$ denote the set of solutions of F . Let $f = f(n)$ be an integer between 1 and n . Construct a graph $G(\Phi(F), f)$ as follows. The set of vertices in $G(\Phi(F), f)$ is $\Phi(F)$; two vertices x and y are adjacent if their Hamming distance is at most f . Now, we say two solutions σ_1 and σ_2 of F are *f-connected* if σ_1 and σ_2 are in the same component of $G(\Phi(F), f)$; conversely, we say σ_1 and σ_2 are *f-separated* if they are not *f-connected*. Given a set of solutions S , we say S is *f-connected* if all solutions in S are in the same component in $G(\Phi(F), f)$; i.e. all solutions in S are pairwise *f-connected*. Given two disjoint sets of solutions S_1 and S_2 , we say they are *f-separated* if for every $\sigma_1 \in S_1$, $\sigma_2 \in S_2$, σ_1 and σ_2 are *f-separated*.

The clustering phenomenon described in Section 1 basically says that given a particular CSP, there exist a constant $c^* > 0$ and two functions $f(n) = o(n)$ and $g(n) = \Omega(n)$ such that, if the density of a random CSP instance F is below c^* , then a.a.s. all solutions of F are *f(n)*-connected; if the density of F is above c^* , then a.a.s. the solutions can be partitioned into

many clusters, each cluster corresponding to a component in $G(\Phi(F), f(n))$: every cluster is $f(n)$ -connected whereas every pair of clusters are $g(n)$ -separated. This is to say that one can walk from any solution to any other inside the same cluster by changing at most $f(n)$ variables at a time; but to walk from one solution to another one in a different cluster, one must change more than $g(n)$ variables in one step. It has been proved [2] that for a random r -XORSAT whose density is a constant not equal to c_r^* , $f(n)$ can be chosen as $C \log n$ for some sufficiently large constant $C > 0$.

2.3 Solution clusters

In this section, we give a full characterisation of clusters of a random linear equation system (or equivalently r -XORSAT).

Given a hypergraph H , its k -core, denoted by $\mathcal{C}_k(H)$, is the maximum subgraph of H in which every vertex has degree at least k . The k -core of a hypergraph can be easily obtained by removing repeatedly every vertex with degree less than k .

The 2 -core of a system of linear equations is the subset of equations corresponding to the hyperedges that are in the 2-core of the underlying hypergraph. The 2-core of $X_r(n, m)$ plays an essential role of determining the clustering threshold for and characterising the clusters of $X_r(n, m)$. It is easy to see that every solution to the 2-core of $X_r(n, m)$ (corresponding to the 2-core of $\mathcal{H}_r(n, m)$) can be easily extended to a solution of the entire system, by setting the other variables in the reverse order that they are removed.

Roughly speaking, the clusters correspond to solutions of the 2-core. But this is not quite true - we need to account for the effects of short *flippable cycles* which we define as follows:

Definition 1. A *flippable cycle* in a hypergraph H is a set of vertices $S = \{v_0, \dots, v_t\}$ where the set of hyperedges incident to S can be ordered as e_0, \dots, e_t such that each vertex v_i lies in e_i and in e_{i+1} and in no other edges of H (addition mod t).

Thus, the vertices v_0, \dots, v_t must have degree exactly two in the hypergraph. The remaining vertices in hyperedges e_0, \dots, e_t can have arbitrary degree and are *not* part of the flippable cycle.

Definition 2. A *core flippable cycle* in a hypergraph H is a flippable cycle in the subhypergraph induced by the 2-core of H .

Thus, in a core flippable cycle, the vertices v_0, \dots, v_t have degree exactly two *in the 2-core*, but possibly higher degree in H . Note also that H may contain flippable cycles outside the 2-core. If we take a solution σ to the entire system, and change the assignment to each variable in a flippable cycle of the underlying hypergraph, we obtain another solution σ' . If the change is on a small core flippable cycle, then σ and σ' differ by a small number of variables and thus they should be in the same cluster, even though they do not agree on the set of variables in the 2-core. This suggests that we have to take core flippable cycles into consideration when we characterise the structure of the clusters.

In Section 5, we will show that for the random hypergraphs studied in this paper, very few vertices lie in core flippable cycles.

Definition 3. Two solutions are *cycle-equivalent* if on the 2-core they differ only on variables in core flippable cycles (while they may differ arbitrarily on variables not in the 2-core).

Definition 4. The *solution clusters* of $X_r(n, m)$ are the cycle-equivalence classes, i.e., two solutions are in the same cluster iff they are cycle-equivalent.

In other words: Let σ be any solution to the subsystem induced by the 2-core. It is easy to see that σ can be extended to a solution to the entire system of equations. All such extensions, along with all extensions of any 2-core solutions obtained by altering σ only on core flippable cycles, form a cluster. By symmetry, all clusters are isomorphic. Note that, if the 2-core is empty, then our definitions imply that all solutions are in the same cluster. So the clustering threshold for $X_r(n, cn)$ corresponds to the emergence threshold of a non-empty 2-core for $\mathcal{H}_r(n, cn)$.

The threshold for the appearance of a non-empty k -core in $\mathcal{H}_r(n, cn)$ ($(k, r) \neq (2, 2)$) is determined [46, 40, 29], given as below.

$$c_{r,k} = \inf_{\mu > 0} \frac{\mu}{r \left[e^{-\mu} \sum_{i=k-1}^{\infty} \mu^i / i! \right]^{r-1}} . \quad (1)$$

Define

$$c_r^* = c_{r,2}. \quad (2)$$

The following theorem confirms that the clustering threshold for $X_r(n, cn)$ is c_r^* .

Theorem 5 ([2, 27]). *For every fixed integer $r \geq 3$ and real number $\epsilon > 0$,*

- (a) *if $c \leq c_r^* - \epsilon$ then all solutions of $X_r(n, cn)$ are $O(\log n)$ -connected;*
- (b) *if $c \geq c_r^* + \epsilon$ then the solutions of $X_r(n, cn)$ are partitioned into well-connected well-separated clusters: every cluster is $O(\log n)$ -connected and every pair of clusters are $\Omega(n)$ -separated.*

2.4 Our contribution

Recall the definition of solution clusters of $X_r(n, cn)$ in Definition 4 and the clustering threshold in (2). We prove that when $c \rightarrow c_r^*$, the connectivity of clusters of $X_r(n, cn)$ undergoes a smooth transition. In the following theorem, we describe the solution geometry of $X_r(n, cn)$ for $c = c_r^* + n^{-\delta}$ when $\delta > 0$ is sufficiently small.

Theorem 6. *Fix constant $r \geq 3$. There exist constants $\kappa = \kappa(r)$, $Z = Z(r) > 0$ such that for any sufficiently small constant $\delta > 0$, in $X_r(n, cn)$ where $c = c_r^* + n^{-\delta}$, a.a.s.:*

- (a) *every cluster is $n^{\kappa\delta}$ -connected;*
- (b) *every pair of clusters are $Zn^{1-r\delta}$ -separated;*
- (c) *there exists a pair of solutions σ, τ in the same cluster such that σ and τ are $n^{\delta/20}$ -separated.*

Remarks.

(i) Theorem 6(b) does not exclude the possibility that different clusters are, in fact, linearly separated; i.e. that there is some constant $\alpha > 0$ such that a.a.s. every pair of clusters are αn -separated.

(ii) In fact, we prove that Theorem 6(a,b) holds for all $\delta < \frac{1}{2}$, but note that these results are trivial for $\delta \geq \frac{1}{\kappa}$ (part (a)) and $\delta \geq \frac{1}{r}$ (part (b)).

(iii) Theorem 6(c) shows that Theorem 6(a) is best possible, up to the value of κ .

For $c = c_r^* - n^{-\delta}$, we prove that all solutions are contained in a single cluster that is $n^{O(\delta)}$ -connected.

Theorem 7. *For $r \geq 3$, there exists $\kappa = \kappa(r)$ such that: for any $0 < \delta < 1/2$ and $c = c_r^* - n^{-\delta}$, a.a.s. all solutions of $X_r(n, cn)$ are $n^{\kappa\delta}$ -connected.*

We conjecture that the above theorem is tight up to the value of κ , just like Theorem 6(a) is. However, the proof of Theorem 6(c) does not generalise to $c = c_r^* - n^{-\delta}$.

Conjecture 8. *For $r \geq 3$, there exists $\kappa' = \kappa'(r)$ such that: for any sufficiently small $\delta > 0$ and $c = c_r^* - n^{-\delta}$, a.a.s. there exist two solutions σ, τ of $X_r(n, cn)$ such that σ and τ are $n^{\kappa'\delta}$ -separated.*

3 Size of the 2-core

Our analysis relies on the size of the 2-core of $\mathcal{H}_r(n, cn)$. This (indeed the k -core for every $k \geq 2$) has been well studied in [46, 40, 29]. Define

$$f_k(\lambda) = e^{-\lambda} \sum_{i \geq k} \frac{\lambda^i}{i!};$$

$$h(\mu) = h_{r,k}(\mu) = \frac{\mu}{f_k(\mu)^{r-1}}.$$

Note that $f_k(\lambda)$ is the probability that a Poisson variable with mean λ is at least k . Fix $r, k \geq 2, (r, k) \neq (2, 2)$; for any $c \geq c_{r,k}$, we define:

$\mu(c)$ is the larger solution to $c = h(\mu)/r$; and

$$\alpha(c) = f_k(\mu(c)), \quad \beta(c) = \frac{1}{r} \mu(c) f_{k-1}(\mu(c)).$$

Define

$$\mu_{r,k} = \mu(c_{r,k}), \quad \alpha_{r,k} = f_k(\mu_{r,k}), \quad \beta_{r,k} = \frac{1}{r} \mu_{r,k} f_{k-1}(\mu_{r,k}). \quad (3)$$

For ease of notation, we drop most of the r, k subscripts.

We will use the following result by Kim [29] (for $k = 2$).

Theorem 9. *Fix $r, k \geq 2, (r, k) \neq (2, 2)$ and an arbitrary constant $\epsilon > 0$.*

(a) For $c \leq c_{r,k} - n^{-1/2+\epsilon}$, a.a.s. the k -core of $\mathcal{H}_r(n, cn)$ is empty.

- (b) For $c \geq c_{r,k} + n^{-1/2+\epsilon}$, a.a.s. $\mathcal{H}_r(n, cn)$ has a k -core with $\alpha(c)n + O(n^{3/4})$ vertices and $\beta(c)n + O(n^{3/4})$ hyperedges.

With some elementary calculations on $\alpha(c)$ and $\beta(c)$ when $c = c_{r,k} + o(1)$, the following lemma follows immediately from Theorem 9. The detailed calculations can be found in [24, Lemma 8].

Lemma 10. *For $r, k \geq 2$, $(r, k) \neq (2, 2)$, there exist three positive constants $K_1 = K_1(r, k)$, $K_2 = K_2(r, k)$, $K_3 = K_3(r, k)$ such that: if $c = c_{r,k} + n^{-\delta}$ for some constant $0 < \delta < 1/2$, then a.a.s. the k -core of $\mathcal{H}_r(n, cn)$ has $\alpha n + K_1 n^{1-\delta/2} + O(n^{1-\delta} + n^{3/4})$ vertices, $\beta n + K_2 n^{1-\delta/2} + O(n^{1-\delta} + n^{3/4})$ hyperedges, and average degree $r\beta/\alpha + K_3 n^{-\delta/2} + O(n^{-\delta} + n^{-1/4})$.*

4 AP-model and degree distribution of the k -core

As usual, we will analyze the k -core using a model which allows us to capture the degree sequence. We will use the AP-model (the allocation and partitioning model), first introduced in [7].

We start with n distinct bins and rm vertex-copies. The probability space generated by the AP-model, denoted by $AP_r(n, m)$ can be described as follows: Allocate each vertex-copy uniformly at random to one of the n bins; take a uniform partition of the rm vertex-copies into m parts; each of size r . The resulting is a configuration, which is a random element in $AP_r(n, m)$. Representing each bin as a vertex and each r -tuple in the partition as a hyperedge, a configuration in $AP_r(n, m)$ corresponds to a multihypergraph on n vertices and m hyperedges.

Of course, a configuration in $AP_r(n, m)$ does not necessarily correspond to a simple hypergraph, but an easy counting argument shows that every simple hypergraph in $\mathcal{H}_r(n, m)$ corresponds to the same number of configurations in $AP_r(n, m)$ and thus, $\mathcal{H}_r(n, m)$ is the random hypergraph generated by $AP_r(n, m)$, conditional on being simple. When $m = O(n)$, the probability that $AP_r(n, m)$ generates a simple hypergraph is $\Theta(1)$ [9]. Thus we immediately have the following corollary.

Corollary 11. *For any $m = O(n)$: If A_n is an event that a.a.s. holds in $AP_r(n, m)$, then A_n holds a.a.s. in $\mathcal{H}_r(n, m)$.*

Note that all bounds from Section 3 on the size of the 2-core hold for the AP-model (indeed these bounds can be obtained by analysing the AP-model). We will also use the following result on the degree sequence of the k -core. A proof can be found in [7, Corollary 2].

Proposition 12. *Let ζ denote the average degree of the k -core of $AP_r(n, cn)$. For any constant $j \geq k$, let ρ_j denote the proportion of vertices in the k -core whose degree equals j . Then, for any $\epsilon > 0$, a.a.s.*

$$\rho_j = e^{-\lambda} \frac{\lambda^j}{f_k(\lambda)j!} + O(n^{-1/2+\epsilon}), \quad (4)$$

where λ satisfies $\lambda f_{k-1}(\lambda)/f_k(\lambda) = \zeta$.

When c is very close to $c_{r,k}$, many parameters in the k -core of $AP_r(n, cn)$ (if it is not empty) are very close to certain critical values. For instance, it is easy to see that the average degree of the k -core is very close to

$$\zeta := \frac{r\beta}{\alpha} = \mu_{r,k} \frac{f_{k-1}(\mu_{r,k})}{f_k(\mu_{r,k})}.$$

The following equation has appeared in several prior papers relating to the analysis the k -core. For instance, it is displayed in [24, Eq. (11)]. We will use this equation in Section 5.

$$e^{-\mu_{r,k}} \frac{\mu_{r,k}^{k-1}}{(k-2)! f_{k-1}(\mu_{r,k})} = \frac{1}{(r-1)}. \quad (5)$$

5 Bounding core flippable cycles

As core flippable cycles influence the clusters of $X_r(n, cn)$, we prove in this section that not many vertices lie on core flippable cycles.

Lemma 13. *Fix $r \geq 3$. For any $0 < \delta < \frac{1}{2}$ and $c = c_{r,2} + n^{-\delta}$, a.a.s. the total sizes of all core flippable cycles in $\mathcal{H}_r(n, cn)$ is at most $O(n^{\delta/2} \log n)$.*

Proof We follow a similar analysis to that in [2, Lemma 35]. We work in the AP-model; Corollary 11 then implies that the result holds for $\mathcal{H}_r(n, cn)$. Recall the definitions of $\mu(c)$, $\alpha(c)$ and $\beta(c)$ from the beginning of Section 3. By Theorem 9, a.a.s. the 2-core contains Q vertices with total degree Λ where

$$Q = \alpha(c)n + O(n^{3/4}), \quad \Lambda = r\beta(c)n + O(n^{3/4}). \quad (6)$$

Let Q_2 denote the number of vertices in the 2-core with degree 2. We will prove below that

$$\frac{2(r-1)Q_2}{\Lambda} \leq 1 - Kn^{-\delta/2}, \quad (7)$$

for some constant $K > 0$.

For any $a \geq 1$, we let X_a denote the number of core flippable cycles of size a . The calculations from the proof of Lemma 35 of [2], which are in fact a simple exercise, say:

$$\mathbb{E}X_a \leq \binom{Q_2}{a} \frac{(a-1)!}{2} 2^a \prod_{\ell=1}^a \frac{r-1}{\Lambda - 2\ell + 1} < \frac{1}{2a} \prod_{\ell=1}^a \frac{(r-1)(2Q_2 - 2\ell + 2)}{\Lambda - 2\ell + 1}.$$

Since $r-1 \geq 2$, we have $2Q_2/\Lambda \leq \frac{1}{2}$ and so $\frac{2Q_2 - 2\ell + 2}{\Lambda - 2\ell + 1} \leq \frac{2Q_2}{\Lambda}$ for each ℓ . So (7) yields:

$$\mathbb{E}X_a \leq \frac{1}{2a} \prod_{\ell=1}^a \frac{2(r-1)Q_2}{\Lambda} \leq \frac{1}{2a} (1 - Kn^{-\delta/2})^a. \quad (8)$$

Now let X denote the total number of vertices appearing on core flippable cycles.

$$\mathbb{E}X \leq \mathbb{E} \sum_{a \geq 1} aX_a < \sum_{a \geq 1} (1 - Kn^{-\delta/2})^a < \frac{1}{K}n^{\delta/2}.$$

So Markov's Inequality yields $\mathbf{Pr}(X > n^{\delta/2} \log n) = o(1)$ which proves the lemma.

It only remains to prove (7). By Proposition 12, for any $\epsilon > 0$, a.a.s.

$$\frac{Q_2}{Q} = e^{-\lambda} \frac{\lambda^2}{2f_2(\lambda)} + O(n^{-1/2+\epsilon}),$$

where λ satisfies $\lambda f_1(\lambda)/f_2(\lambda) = \Lambda/Q$. Conditional on any values of Q and Λ satisfying (6), we have

$$\lambda \frac{f_1(\lambda)}{f_2(\lambda)} = \frac{r\beta(c)}{\alpha(c)} + O(n^{-1/4}).$$

The function $g(x) = xf_1(x)/f_2(x)$ is a strictly increasing function on $x > 0$ (see [24, Lemma 26] for a proof). Then, by the definition of $\alpha(c)$, $\beta(c)$ and $\mu(c)$ above (3), we have

$$\lambda = \mu(c) + O(n^{-1/4}).$$

Immediately, we have

$$Q_2 = \left(\frac{e^{-\mu(c)}\mu(c)^2}{2f_2(\mu(c))} + O(n^{-1/4}) \right) Q = \left(\frac{e^{-\mu(c)}\mu(c)^2}{2f_2(\mu(c))} + O(n^{-1/4}) \right) \alpha(c)n, \quad (9)$$

since the function $e^{-\lambda}\lambda^2/2f_2(\lambda)$ has bounded derivative at $\lambda = \mu(c)$ and the error $O(n^{-1/2+\epsilon})$ is absorbed by $O(n^{-1/4})$. By (5),

$$e^{-\mu_{r,2}} \frac{\mu_{r,2}}{f_1(\mu_{r,2})} = \frac{1}{(r-1)}.$$

It is easy to check that the derivative of $e^{-\mu} \frac{\mu}{f_1(\mu)}$ with respect to μ is strictly negative in a small neighbourhood of $\mu_{r,2}$. By Lemma 10, $\mu(c) = \mu_{r,2} + K_2 n^{-\delta/2} + o(n^{-\delta/2})$ for some constant $K_2 > 0$. Hence,

$$e^{-\mu(c)} \frac{\mu(c)}{f_1(\mu(c))} = e^{-\mu_{r,2}} \frac{\mu_{r,2}}{f_1(\mu_{r,2})} - K_3 n^{-\delta/2} + o(n^{-\delta/2}) = \frac{1}{(r-1)} - K_3 n^{-\delta/2} + o(n^{-\delta/2}), \quad (10)$$

for some constant $K_3 > 0$. Now, by (6) and (9), and recalling the definition of $\alpha(c)$ and $\beta(c)$ above (3),

$$\begin{aligned} \frac{2(r-1)Q_2}{\Lambda} &= \left(e^{-\mu(c)} \frac{\mu(c)^2}{2f_2(\mu(c))} + O(n^{-1/4}) \right) \frac{2(r-1)\alpha(c)}{r\beta(c)} \\ &= \left(e^{-\mu(c)} \frac{\mu(c)^2}{2f_2(\mu(c))} + O(n^{-1/4}) \right) 2(r-1) \frac{f_2(\mu(c))}{\mu(c)f_1(\mu(c))} \\ &= (r-1) \cdot e^{-\mu(c)} \frac{\mu(c)}{f_1(\mu(c))} + O(n^{-1/4}) \\ &= 1 - K_3(r-1)n^{-\delta/2} + O(n^{-1/4}) + o(n^{-\delta/2}) \quad \text{by (10).} \end{aligned}$$

Since $\delta < 1/2$, we have $n^{-1/4} = o(n^{-\delta/2})$. Then (7) follows. \square

Next we show that w.h.p. the core flippable cycles are disjoint:

Lemma 14. Fix $r \geq 3$. For any $0 < \delta < \frac{1}{2}$ and $c = c_{r,2} + n^{-\delta}$, a.a.s. no vertex lies in two different core flippable cycles.

Proof If a vertex v lies in two core flippable cycles, then we must have the following structure (see a detailed description in the proof of Lemma 35 in the arXived version of [2], which uses different notation): One core flippable cycle with vertices $v = v_1, \dots, v_a$ where each pair $v_i, v_{i+1}, i = 1, \dots, a$ shares a hyperedge; and a sequence of vertices $u_1, \dots, u_q, q \leq a$ (they form part of the other core flippable cycle) where (i) each pair $u_i, u_{i+1}, i = 1, \dots, q - 1$ shares a hyperedge which contains none of v_1, \dots, v_a , (ii) u_1 is in the hyperedge containing v_j, v_{j+1} and (iii) u_q is in the hyperedge containing $v_{j'}, v_{j'+1}$ for some $j' \neq j$. All these vertices have degree two.

Calculations very similar to those in the previous proof bound the expected number of such structures for a given value of $a = o(n)$ as follows. The a^2 term comes from a choices for each of j, j', q multiplied by the $\frac{1}{2a}$ term from (8); the $\frac{1}{\Lambda}$ term comes from the fact that the double cycle produces $a + q + 1$ different $\Theta(\frac{1}{\Lambda})$ terms but only $a + q$ different $\Theta(n) = \Theta(\Lambda)$ terms.

$$O\left(\frac{a^2}{\Lambda}\right) (1 - Kn^{-\delta/2})^{a+q}.$$

Lemma 13 allows us to restrict to $a = O(n^{\delta/2} \log n)$. Summing over all such a yields that the expected number of these structures is at most

$$O(1) \frac{(n^{\delta/2} \log n)^3}{\Lambda} (1 - Kn^{-\delta/2})^a = O(n^{3\delta/2-1} \log^3 n) = o(1),$$

for $\delta < \frac{1}{2}$. □

6 Proof of Theorems 6(a) and 7

Recall that the k -core of a hypergraph H can be obtained by repeatedly removing vertices with degree less than k .

Definition 15. A k -stripping sequence is a sequence of vertices that can be deleted from a hypergraph, one-at-a-time, along with their incident hyperedges such that at the time of deletion each vertex has degree less than k .

Let H be an r -uniform hypergraph and let $\Psi = v_1 v_2 \dots$ be a k -stripping sequence which contains all non- k -core vertices of H . We create a directed graph (not a directed hypergraph) $\mathcal{D}(\Psi)$ associated with Ψ as follows. The vertices in $\mathcal{D}(\Psi)$ are a subset of the vertices of H – specifically, the vertices not in the k -core and the vertices of the k -core that have a neighbour outside of the k -core. At the moment when v_i is to be deleted from H , consider each hyperedge x that is incident with v_i (there are at most $k - 1$ of them). Add a directed edge to v_i in $\mathcal{D}(\Psi)$ from each of the $r - 1$ other vertices in x .

For any vertex $v \in \mathcal{D}(\Psi)$, we define $R_{\Psi}^+(v)$ to be the set of vertices reachable from v in $\mathcal{D}(\Psi)$. We have the following bound on $|R_{\Psi}^+(v)|$. Part (a) is from [24, Theorem 43] (for $c = c_{r,k} + n^{-\delta}$) and [25, Theorem 6] (for $c = c_{r,k} - n^{-\delta}$), and part (b) is from [24, Theorem 5] (for $c = c_{r,k} + n^{-\delta}$) and [25, Theorem 4] (for $c = c_{r,k} - n^{-\delta}$).

Theorem 16. *Let $r, k \geq 2$, $(r, k) \neq (2, 2)$ be fixed. There is a constant $\kappa > 0$ such that: for any constant $0 < \delta < \frac{1}{2}$, if $c = c_{r,k} \pm n^{-\delta}$, then*

- (a) *a.a.s. there is a k -stripping sequence Ψ containing all non- k -core vertices of $\mathcal{H}_r(n, cn)$ such that $|R_{\Psi}^+(v)| \leq n^{\kappa\delta}$ for all $v \in \Psi$.*
- (b) *a.a.s. for every k -stripping sequence Ψ , there exists a non- k -core vertex v for which $|R_{\Psi}^+(v)| = \Omega(n^{\delta/2})$.*

For the purposes of studying r -XORSAT clusters, we will only apply Theorem 16 for $k = 2$. We use the stripping sequence Ψ guaranteed by Theorem 16(a).

The argument that our upper bound on $|R_{\Psi}^+(v)|$ implies that the clusters are well-connected is the same as that used in [2]. We include it here for exposition, and because it is needed to understand the proof of Theorem 6(c).

Each cluster is specified by an assignment to all of the 2-core variables not in any core flippable cycles; the cluster consists of all extensions from that assignment to the remaining variables. More specifically, choose any such assignment σ which does not violate any equations, substitute the value $v = \sigma(v)$ for each v in the range of σ . This removes some of the linear equations (those whose variables are all set) and removes variables from some others. The cluster now consists of all solutions to this reduced system.

If we apply Gaussian elimination to the equations of the reduced system, beginning with those in the core flippable cycles and then proceeding in the reverse order in which the equations were removed by Ψ , then we will obtain a system of equations in which each variable is expressed as the sum of a subset of what we call \mathcal{F} , the *free variables*. It is a simple exercise (recall Lemma 14 and see [2] for more details) to show that the free variables are as follows:

Observation 17. \mathcal{F} consists of the variables corresponding to: (i) the non-2-core vertices with indegree zero in $\mathcal{D}(\Psi)$; (ii) one vertex from each core flippable cycle.

For each variable $v \notin \mathcal{F}$ of the reduced system, there will be a set of free variables $\chi(v) \subseteq \mathcal{F}$, such that the Gaussian elimination leaves exactly one equation containing v , and it is of the form

$$v = z_v + \sum_{u \in \chi(v)} u, \quad (11)$$

where z_v is determined by σ and thus is fixed for each cluster. Furthermore, each variable of $\chi(v)$ can reach v in $\mathcal{D}(\Psi)$. (Note that $\chi(v)$ does not necessarily contain every free variable that can reach v in $\mathcal{D}(\Psi)$.) Every equation is of this form, where $v \notin \mathcal{F}$. So for each $u \in \mathcal{F}$ we set $\chi(u) = \{u\}$ and $z_u = 0$. Each of the $2^{|\mathcal{F}|}$ possible assignments to the free variables is permissible.

It is important to stress that the set of free variables, and furthermore the sets $\chi(v)$, are determined only by our application of Gaussian elimination to the hypergraph, and not by any particular solution. So these are the same for every cluster.

Each cluster contains exactly $2^{|\mathcal{F}|}$ solutions, one for each assignment to the free variables. We can move from any solution to any other solution by changing the free variables, one-at-a-time, and updating the non-free variables using (11). In the step where we change the

free variable u , we only change the variables in

$$\chi^{-1}(u) = \{v : u \in \chi(v)\}.$$

If u is not in the 2-core, then $\chi^{-1}(u) \subseteq R_{\Psi}^{+}(u)$ and so Theorem 16(a) implies that we change at most $n^{\kappa\delta}$ variables. If u is a free variable on a core flippable cycle C , then $\chi^{-1}(u)$ is contained in $\cup_w R_{\Psi}^{+}(w)$ over all $w \in \mathcal{D}(\Psi) \cap C$. (Recall that the core flippable cycles are disjoint by Lemma 14.) It is easy to show that a.a.s. the maximum degree of $\mathcal{H}_r(n, cn)$ is less than $\log n$, and so Theorem 16(a) implies $|R_{\Psi}^{+}(w)| \leq n^{\kappa\delta} r \log n$ for each w in a core flippable cycle. By Lemma 13, a.a.s. the number of vertices in core flippable cycles is at most $n^{\delta/2} \log n$ and so $|\chi^{-1}(u)| \leq n^{\kappa\delta} r \log n \times n^{\delta/2} \log n < n^{(\kappa+1)\delta}$. This proves Theorems 6(a) and 7. \square

7 Proof of Theorem 6(b)

In this section, we describe the proof that solutions in different clusters are $\Theta(n^{1-r\delta})$ -separated. As the proof is very similar to that of [2, Theorem 2], we only sketch the differences.

Proof of Theorem 6(b) (sketch). This follows the same argument as the proof of Theorem 2 of [2]. The only change is to Lemma 51 of [2], where instead of proving that a.a.s. there is no non-empty linked set (see [2] for definitions) of size less than αn , we prove that a.a.s. there is none of size less than $n^{1-r\delta}$. (Caution: in [2] the usage of k, r is inverted from that of this paper.)

As in [2], we use X_a to denote the number of linked sets S with $|\Gamma(S)| = a$ (see the definition of $\Gamma(S)$ in [2]). Property (iii) at the beginning of the proof of Lemma 51, is equivalent to saying $\frac{2(r-1)Q_2}{\Lambda} \leq 1 - \zeta$, for some $\zeta > 0$, in the notation of this paper. Instead, we have $\frac{2(r-1)Q_2}{\Lambda} \leq 1 - Kn^{-\delta/2}$ by (7). This results in replacing $Z_1 = \Theta(1)$ from the proof of Lemma 51 of [2] with $Z_1 = \Theta(n^{\delta/2})$ (in the notation of [2]); this yields

$$\mathbb{E}(X_a) < \left(\frac{\Theta(an^{r\delta})}{n} \right)^{a/2r},$$

and it then follows easily that $\mathbb{E} \left(\sum_{a=1}^{Zn^{1-r\delta}} X_a \right) = o(1)$ for sufficiently small $Z = Z(r) > 0$. This proves the theorem. \square

8 Proof of Theorem 6(c)

We will make use of the characterization of the solution space in terms of free variables introduced in Section 6. We will define a 2-stripping sequence Ψ , which in turn will define a set \mathcal{F} of free variables. Recalling that each cluster contains one solution for each of the $2^{|\mathcal{F}|}$ settings of the variables of \mathcal{F} , it will suffice to show that there is at least one free variable u such that any step which changes the value of u will also change at least $n^{\delta/20}$ other variables.

Recall that if we change a free variable u , and no other free variables, then we also change all the variables in $\chi^{-1}(u) \subseteq R_{\Psi}^{+}(u)$. Theorem 16 says that there is at least one variable u

with $|R_\Psi^+(u)| \geq n^{\Theta(\delta)}$. However, this does not immediately imply that each solution cluster is not $n^{\Theta(\delta)}$ -connected. For one thing, we require that there is a *free* variable u such that $R_\Psi^+(u)$ is that large. In fact, we actually require $\chi^{-1}(u) \subseteq R_\Psi^+(u)$ to be that large. But even that would not suffice, since it would only imply that a step where u is the only free variable changed would require changing $n^{\Theta(\delta)}$ other variables. It still leaves open the possibility that one could change u using a step that also changes other free variables w_1, \dots, w_t where $\cup \chi^{-1}(w_i)$ intersects $\chi^{-1}(u)$ and so not every variable in $\chi^{-1}(u)$ is changed.

To prove Theorem 6(c) we prove that for our choice of Ψ , we have:

Property 18. *There is a free variable u , along with $n^{\delta/20}$ other variables $v_1, \dots, v_{n^{\delta/20}}$ such that for each i , $\chi(v_i) = \{u\}$.*

In order to move through every solution in a cluster, eventually there must be a step where u is changed. At that step, each of $v_1, \dots, v_{n^{\delta/20}}$ are changed as well, regardless of which other free variables are changed. So the cluster is not $n^{\delta/20}$ -connected.

In order to find such u and $v_1, \dots, v_{n^{\delta/20}}$, we need to specify a stripping sequence Ψ . We start with defining a parallel stripping process which produces the k -core.

Definition 19. The *parallel k -stripping process*, applied to a hypergraph H , consists of iteratively removing *all* vertices of degree less than k at once along with any hyperedges containing any of those vertices, until no vertices of degree less than k remain. Let S_i denote the set of vertices that are removed during iteration i . We use \hat{H}_i to denote the hypergraph remaining after $i - 1$ iterations, i.e. after removing S_1, \dots, S_{i-1} .

Note that the parallel k -stripping process terminates with the k -core of H . In order to define Ψ , we consider a slowed-down version of the parallel stripping process, called SLOW-STRIP. To maintain SLOW-STRIP, we use a queue \mathcal{Q} . Initially, \mathcal{Q} is the set of all light vertices (vertices with degree less than k) of H . In each step of SLOW-STRIP, a hyperedge x incident with the light vertex in the front of \mathcal{Q} is removed. If any vertex becomes light after the removal of x , add it to the end of \mathcal{Q} . Remove the light vertex in the front of \mathcal{Q} if its degree drops to zero.

Let Ψ be the stripping sequence produced by SLOW-STRIP and let $\mathcal{D} = \mathcal{D}(\Psi)$ be the digraph associated with Ψ (recall the definition of \mathcal{D} below Definition 15). Let I^* be the largest value of i such that S_i contains a vertex with indegree zero in \mathcal{D} (recall that a vertex with indegree zero in \mathcal{D} is a free variable), and let $u^* \in S_{I^*}$ be some such vertex.

Our first step is to prove that there are no other free vertices within $n^{\delta/20}$ levels of u^* :

Lemma 20. *Assume that $\delta > 0$ is sufficiently small. A.a.s u^* is the only free vertex in $\cup_{i \geq I^* - n^{\delta/20}} S_i$.*

Given a non-core vertex w , we define $T(w)$ to be the set of vertices v that can reach w in \mathcal{D} ; i.e. the set of vertices v such that $w \in R_\Psi^+(v)$. For $u \in T(w)$, define $T(w, u)$ to be the subgraph of $T(w)$ containing all vertices reachable from u ; i.e. vertices on walks from u to w . We will prove:

Lemma 21. *Assume that $\delta > 0$ is sufficiently small. A.a.s. there is some $w \in S_{I^* - n^{\delta/20}}$ such that*

- (a) $u^* \in T(w)$;
- (b) $|T(w)| \leq n^{3\delta/2}$;
- (c) the subgraph of \mathcal{D} induced by the vertices of $T(w, u^*)$ is a directed path, containing exactly one vertex in each S_i , $I^* - n^{\delta/20} \leq i \leq I^*$.

We defer the proofs of Lemmas 20 and 21 to Section 8.2.

Proof of Theorem 6(c). Choose a vertex $w \in S_{I^* - n^{\delta/20}}$ in \mathcal{H} satisfying Lemma 21. Our strategy will be to show that Property 18 holds for u^* and the $n^{\delta/20}$ variables on the directed path $T(w, u^*)$. Thus we want to show that for each vertex v on this path, $\chi(v)$ contains no free vertices other than u^* . Lemma 20 will establish that there are no such free vertices outside of the 2-core. So we will begin by arguing that a.a.s. each $\chi(v)$ contains no free vertices inside the 2-core; i.e. the free vertices on core flippable cycles.

By Lemma 21(b) at most $n^{3\delta/2}$ vertices in $\mathcal{C}_2 = \mathcal{C}_2(\mathcal{H})$ are adjacent to $T(w) \setminus \mathcal{C}_2$. We first prove that a.a.s. none of these vertices in \mathcal{C}_2 are contained in a core flippable cycle of \mathcal{H} , and thus a.a.s. no vertex of any core flippable cycle is in $T(w)$.

We can choose \mathcal{H} in the following way. First, choose a random hypergraph $\mathcal{H}_1 = \mathcal{H}_r(n, cn)$. Then form \mathcal{H}_2 by randomly permuting the vertices of the 2-core of \mathcal{H}_1 . That is: let σ be a uniformly random permutation of the vertices of the 2-core. Replace every hyperedge (v_1, \dots, v_r) in the 2-core with $(\sigma(v_1), \dots, \sigma(v_r))$, and keep every hyperedge with at least one non-2-core vertex unchanged. Note that the vertex set of $\mathcal{C}_2(\mathcal{H}_2)$ is equal to the vertex set of $\mathcal{C}_2(\mathcal{H}_1)$; to see this, consider any stripping sequence which when produces the 2-core of \mathcal{H}_1 - the same sequence will produce the same set of vertices as the 2-core of \mathcal{H}_2 .

We claim that \mathcal{H}_2 is distributed like $\mathcal{H}_r(n, cn)$ and hence is a valid choice for \mathcal{H} . To see this, partition the set of hypergraphs with cn edges into equivalence classes where $\mathcal{H} \sim \mathcal{H}'$ if $\mathcal{C}_2(\mathcal{H}), \mathcal{C}_2(\mathcal{H}')$ are isomorphic and have the same vertex set. The procedure in the previous paragraph first chooses an equivalence class with probability proportional to its size, and then chooses a uniform member of that class. Thus it picks a uniform hypergraph with cn hyperedges.

As we said above, at most $n^{3\delta/2}$ vertices in $\mathcal{C}_2(\mathcal{H}_1)$ are adjacent to $T(w) \setminus \mathcal{C}_2$. By Lemma 13, the total number of vertices contained in core flippable cycles of $\mathcal{C}_2(\mathcal{H}_1)$ is $O(n^{\delta/2} \log^2 n)$. So after taking the random permutation, the probability that any of the vertices adjacent to $T(w) \setminus \mathcal{C}_2$ is contained in a core flippable cycle of \mathcal{H}_2 is $O(n^{3\delta/2 + \delta/2 - 1} \log^2 n) = o(1)$, as $\delta < 1/2$. This confirms that a.a.s. $T(w)$ contains no vertex of a core flippable cycle of $\mathcal{H}_r(n, cn)$.

Now consider any vertex $v \in T(w, u^*)$. Since $T(w, u^*)$ induces a directed path in \mathcal{D} by Lemma 21(c), a.a.s. there is exactly one directed path from u^* to v , and it follows easily that $u^* \in \chi(v)$. Since $T(v) \subseteq T(w)$, a.a.s. no vertex of any core flippable cycle is in $T(v)$, as a.a.s. none is in $T(w)$. By Lemma 20, u^* is the only vertex in $\cup_{i \geq I^* - n^{\delta/20}} S_i$ with indegree zero in \mathcal{D} . Therefore, u^* is the only free variable in $T(v)$ and so $\chi(v) = \{u^*\}$, as we wanted to prove.

Therefore, if any two solutions in the same cluster differ on u^* then they differ on all of the $n^{\delta/20} + 1$ variables on the path from u^* to w by Lemma 21(c). Since every cluster contains one solution for each setting of the free variables, this implies that every cluster is not $n^{\delta/20}$ -connected. This completes the proof for Theorem 6(c). \square

8.1 Parallel stripping process and SLOW-STRIP

To prove Lemmas 20 and 21, we will work in the AP-model. Corollary 11 then implies that these Lemmas hold for $\mathcal{H}_r(n, cn)$.

In this subsection, we state some results on parallel 2-stripping process including the number of iterations the process takes and the changes of $|S_i|$ in each iteration.

Note that we only need to consider $c = c_r^* + n^{-\delta}$ by the hypotheses of Theorem 6.

The following theorem, from [24, Theorem 3], bounds the number of iterations the parallel 2-stripping sequence takes. The original statement was for $\mathcal{H}_r(n, cn)$ and holds for any k -stripping sequence where $(k, r) \neq (2, 2)$, but the proof used $AP_r(n, cn)$ and Corollary 11.

Theorem 22. *Fix $r \geq 3$. For any constant $0 < \delta < 1/2$ and $c = c_r^* + n^{-\delta}$, a.a.s. the number of iterations the parallel 2-stripping process takes, when it is applied to $AP_r(n, cn)$, is $\Theta(n^{\delta/2} \log n)$.*

Next, we state some properties of $|S_i|$ in the parallel 2-stripping process, applied to $AP_r(n, cn)$, where $c = c_r^* + n^{-\delta}$ and $0 < \delta < 1/2$. These properties follows from [24, Lemma 49].

Lemma 23. *For every $\epsilon > 0$, there exist constants B, Y_1, Y_2, Z_1 , dependent only on r, ϵ , such that a.a.s. $\sum_{i \geq B} |S_i| \leq \epsilon n$, and for every $B \leq i < I_{\max}$ with $|S_i| \geq n^\delta \log^2 n$:*

- (a) *if $|S_i| < n^{1-\delta}$ then $(1 - Y_1 n^{-\delta/2})|S_i| \leq |S_{i+1}| \leq (1 - Y_2 n^{-\delta/2})|S_i|$;*
- (b) *if $|S_i| \geq n^{1-\delta}$ then $\left(1 - Y_1 \sqrt{\frac{|S_i|}{n}}\right) |S_i| \leq |S_{i+1}| \leq \left(1 - Y_2 \sqrt{\frac{|S_i|}{n}}\right) |S_i|$;*
- (c) $\sum_{j \geq i} |S_j| \leq Z_1 |S_i| n^{\delta/2}$;
- (d) *the maximum degree of $AP_r(n, cn)$ is at most $\log n$.*

In each step of SLOW-STRIP on $AP_r(n, cn)$, a vertex-copy of the vertex at the front of \mathcal{Q} is deleted, together with another $r - 1$ vertex-copies that are in the same part (i.e. hyperedge) of the removed vertex-copy. For each i , we define:

$t(i)$ is the step in SLOW-STRIP that the vertex in the front of \mathcal{Q} is the first vertex in S_i to be removed.

I.e. $t(i)$ is the step of SLOW-STRIP corresponding to the beginning of the i th iteration of the parallel stripping process.

We need the following bound on the rate at which vertices become light during SLOW-STRIP, which is not stated explicitly in [24].

Lemma 24. *There is a constant $K > 0$ such that a.a.s. at every step of SLOW-STRIP, the degrees of the remaining vertices are such that the expected number of heavy vertices that become light in that step is at most $1 - Kn^{-\delta/2}$.*

Proof In each such step, we remove a copy of a light vertex and $r - 1$ uniformly chosen copies. The proof of Lemma 16 in [24] establishes that each time we remove a uniformly chosen copy, the probability that a heavy vertex becomes light (i.e. that we remove a copy of a degree two vertex) is at most $1/(r - 1) - \Theta(n^{-\delta/2})$. (See Definition 29, line (38) and the discussion preceding line (33) of [24] and let $k = 2$.) This yields the lemma. \square

8.2 Proof of Lemmas 20 and 21

We will prove that Lemmas 20 and 21 hold in the AP-model. Corollary 11 then shows that they hold for $X_r(n, cn)$.

Proof of Lemma 20. Choose $\nu = \frac{1}{2} - \delta/6$. Run the parallel 2-stripping process and let i_1 denote the first iteration in which $|S_{i_1}| \leq n^\nu$. Our first step will be to show that a.a.s. $I^* > i_1 + n^{\delta/20}$.

Let $i_2 = i_1 + n^{2\delta/5}$ and $i_3 = i_1 + 2n^{2\delta/5}$. By Lemma 23(c), a.a.s. $\sum_{j \geq i_1} |S_j| = O(|S_{i_1}| n^{\delta/2}) = O(n^{\frac{1}{2} + \delta/3}) = o(n^{1-\delta})$ for small δ . Thus, a.a.s. at any iteration $i \geq i_1$, the total number of light vertices (vertices with degree less than 2) is $o(n^{1-\delta})$.

Therefore, we can apply Lemma 23(a) (recursively) to obtain that for all $i_1 \leq i \leq i_3$:

$$|S_i| > (1 - Y_1 n^{-\delta/2})^{2n^{2\delta/5}} |S_{i_1}| = (1 - o(1)) |S_{i_1}| > \frac{1}{2} n^\nu. \quad (12)$$

This is valid since in each recursion $i_1 \leq i \leq i_3$, we have $|S_i| = \Omega(|S_{i_1}|) \geq n^\delta \log^2 n$, provided $\delta < 3/7$, and so the assumption of Lemma 23 is satisfied. Lemma 23(a) also implies that for all $i_1 \leq i \leq i_3$,

$$|S_i| = O(n^\nu). \quad (13)$$

Our proof consists of two steps. First, we show that $I^* \geq i_2$ by proving:

Claim 1: A.a.s. there is a free variable in $\cup_{i_2 \leq i \leq i_3} S_i$.

Then we prove that after level i_1 , the free variables are all separated by many levels:

Claim 2: A.a.s. there is no integer $i > i_1$ such that $\cup_{i'=i}^{i+n^{\delta/20}} S_{i'}$ contains two free variables.

These prove the lemma as follows. Claim 1 implies that $I^* \geq i_2 > i_1 + n^{\delta/20}$. So if there was another level S_i containing a free variable with $i \geq I^* - n^{\delta/20}$, then this would violate Claim 2.

Proof of Claim 2. Recall how SLOW-STRIP is run in the AP-model: Since $k = 2$, every light vertex, i.e. every vertex in \mathcal{Q} , has exactly one copy remaining. At each step, the copy of the vertex at the front of \mathcal{Q} is removed, along with $r - 1$ vertex-copies selected uniformly from amongst all remaining copies; we call each of these $r - 1$ selections a *trial*. A light vertex v becomes a free variable iff its remaining copy is chosen for removal before it reaches the front of \mathcal{Q} . When a light vertex u selects the remaining copy of a light vertex v during one of its $r - 1$ trials, we say that u *frees* v .

We let E_S denote the event that for all $i \geq i_1$, we have $|S_i| \leq n^\nu$; the definition of i_1 and Lemma 23(b) ensure that E_S holds a.a.s.. Note that when the last member of S_i is removed from \mathcal{Q} , then \mathcal{Q} consists all members of S_{i+1} . So \mathcal{Q} can only contain members from two consecutive levels S_i, S_{i+1} . Thus if E_S holds then for all $t \geq t(i_1)$, i.e. for all steps of SLOW-STRIP that correspond to iterations $i \geq i_1$ of the parallel process, we always have $|\mathcal{Q}| < 2n^\nu$.

Now consider any $i > i_1$. If there are two free variables in $\cup_{i'=i}^{i+n^{\delta/20}} S_{i'}$ then there are two different trials in which a vertex is freed by vertices in $\cup_{j=i-1}^{i+n^{\delta/20}} S_j$. If E_S holds then there are at most $(n^{\delta/20} + 2)n^\nu(r - 1)$ such trials and during each such trial, there are at most $2n^\nu$ vertex-copies in \mathcal{Q} , i.e. vertex-copies whose choice would free a vertex. By Theorem 9(b),

we can assume that there are always at least $r\beta(c)n$ remaining vertex-copies. So in each trial, the probability of freeing a vertex is at most $2n^\nu/r\beta(c)n$. Putting this together, the probability that E_S holds and there are two free variables in $\cup_{i'=i}^{i+n^{\delta/20}} S_{i'}$ is at most:

$$\binom{(n^{\delta/20} + 2)n^\nu(r-1)}{2} \left(\frac{2n^\nu}{r\beta(c)n} \right)^2 < n^{-.55\delta},$$

since $\nu = \frac{1}{2} - \delta/6$. Since E_S holds a.a.s., it follows that a.a.s. there is no such i in the range $i_1 \leq i \leq I_{\max}$, where I_{\max} is the last iteration of the parallel stripping process, which is a.a.s. less than $n^{0.51\delta}$ by Theorem 22. This proves the claim.

Proof of Claim 1. Let E_T denote the event that for all $i_2 \leq i \leq i_3$ we have $|S_i| \geq \frac{1}{2}n^\nu$. E_T holds a.a.s. by (12).

If at least one vertex is freed during the removal of the vertices in $\cup_{i_2 \leq i \leq i_3-1} S_i$, then there is a free variable in $\cup_{i_2 \leq i \leq i_3} S_i$. If E_T holds, then the total number of vertices removed is at least $(i_3 - i_2)\frac{1}{2}n^\nu > n^{\nu+2\delta/5}$.

Consider the removal of the first $\frac{1}{2}|S_i|$ vertices of S_i . If no vertices are freed during these removals, i.e. if each time we remove the first member of \mathcal{Q} , we don't select any other member of \mathcal{Q} for deletion, then the size of \mathcal{Q} remains at least $\frac{1}{2}|S_i| > \frac{1}{4}n^\nu$ if E_T holds. If there are no free variables in S_i , then every member of \mathcal{Q} has degree 1 (not 0) and so each deletion of a member of \mathcal{Q} results in $r-1$ trials. At each such trial, the total number of vertex-copies remaining is at most rcn . It follows that the probability that E_T holds and no vertex is freed during the removal of the first $\frac{1}{2}|S_i|$ vertices over all $i_2 \leq i \leq i_3$ is at most

$$\left(1 - \frac{n^\nu/4}{rcn}\right)^{\frac{1}{2}(r-1)n^{\nu+2\delta/5}} = e^{-\Theta(n^{2\delta/5+2\nu-1})} = o(1),$$

since $\nu = \frac{1}{2} - \delta/6$. Because E_T holds a.a.s., this proves Claim 1 and so completes the proof of the lemma. \square

Proof of Lemma 21. As in the previous proof, we let $\nu = \frac{1}{2} - \frac{\delta}{6}$. Run SLOW-STRIP on $AP_r(n, cn)$ and let i_1 denote the first iteration in the parallel stripping process such that $|S_{i_1}| \leq n^\nu$; set $t_1 = t(i_1)$. Recall that \mathcal{C}_2 is the 2-core of $AP_r(n, cn)$. We will prove:

Claim 25. *A.a.s. for every $w \in \cup_{i \geq i_1} S_i$, $T(w) \setminus \mathcal{C}_2$ induces a directed tree rooted at w , where a directed rooted tree means a tree where every edge is directed towards the root.*

Note that every vertex in any S_i must lie in a hyperedge with a vertex in S_{i-1} that was deleted during iteration $i-1$ of the parallel stripping process. Thus each vertex in S_i points to a vertex in S_{i-1} in $\mathcal{D}(\Psi)$. It follows that there is a path in $\mathcal{D}(\Psi)$ from u^* to some vertex $w \in S_{I^*-n^{\delta/20}}$ where that path contains a vertex in each S_i , $I^* - n^{\delta/20} \leq i \leq I^*$. The claim will imply parts (a,c) for this choice of w .

We start by bounding the size of each $T(w) \setminus \mathcal{C}_2$.

Fix some $w \in S_i$, with $i \geq i_1$. We maintain a set $\mathcal{T}(w)$ as follows.

Initially $\mathcal{T}(w) := \{w\}$. Whenever we delete a vertex $v \in \mathcal{Q}$ such that $v \in \mathcal{T}(w)$: (i) each neighbour of v that has degree at most 2, and hence is in or will enter \mathcal{Q} , is placed in $\mathcal{T}(w)$; (ii) each neighbour of v that has degree greater than 2 is coloured Red. Every

time a Red vertex enters \mathcal{Q} , it is placed in $\mathcal{T}(w)$. Thus, when we finish SLOW-STRIP, $\mathcal{T}(w) = T(w) \setminus \mathcal{C}_2$.

We will analyze $\mathcal{T}(w)$ using a branching process. When a vertex $v \in \mathcal{T}(w)$ is deleted by SLOW-STRIP, we say that we are *processing* v . If a vertex u is added to $\mathcal{T}(w)$ while v is being processed then we consider u to be an *offspring* of v . If a vertex u is added to $\mathcal{T}(w)$ during the deletion of a vertex not in $\mathcal{T}(w)$ (and so u must be Red), then we consider u to be an *offspring* of the most recently processed member of $\mathcal{T}(w)$. Note that the offspring of v are not necessarily adjacent to v in \mathcal{D} .

In other words: we say that u is an *offspring* of a vertex $v \in \mathcal{T}(w)$, if u entered $\mathcal{T}(w)$ between the iterations of SLOW-STRIP ranging from the time we remove v up until just before the next iteration where we remove a member of $\mathcal{T}(w)$. Our definition of an offspring may look a little unnatural. This is because we want to include in $\mathcal{T}(w)$ those vertices that become light because of the removal of some vertices not in $\mathcal{T}(w)$, but have already been found to be adjacent to some vertex in $\mathcal{T}(w)$. Note that these vertices are in $T(w) \setminus \mathcal{C}_2$.

There are two scenarios under which u can become an offspring of v : (i) at the time we delete v , u is a neighbour of v and u has degree at most 2; (ii) at the step after v is deleted, u is Red and u enters \mathcal{Q} before the next member of $\mathcal{T}(w)$ is deleted. For case (ii) to occur, u must be the neighbour of another vertex $v' \notin \mathcal{T}(w)$ that is removed from \mathcal{Q} , and the degree of u must drop below 2 when v' is removed.

It will be convenient to consider $\mathcal{T}'(w) \subseteq \mathcal{T}(w)$, which differs from $\mathcal{T}(w)$ in that only the first $n^{2\delta}$ vertices to be coloured Red can enter $\mathcal{T}'(w)$. We will show that, in fact, a.a.s. $\mathcal{T}'(w) = \mathcal{T}(w)$.

When removing $v \in \mathcal{T}'(w)$, the expected number of offspring created under scenario (i) is at most $1 - Kn^{-\delta/2}$ for some constant $K > 0$ by Lemma 24, and this holds for all $t \geq t(i) \geq t(i_1)$.

The number of iterations until the next member of $\mathcal{T}'(w)$ is removed is at most $O(|S_i| \times n^{\delta/2}) = O(n^{\nu+\delta/2})$, by Lemma 23(c). If u is an offspring of v created under scenario (ii), then u must be one of the first $n^{2\delta}$ Red vertices. Furthermore during one of those iterations, exactly two vertex-copies of u remain and one of them is selected. The total number of such vertex-copies over all choices of u is at most $2n^{2\delta}$, and since there are a linear number of vertex-copies to choose from, the expected number of offspring of v created under scenario (ii) is at most

$$O(n^{\nu+\delta/2}) \times O(n^{2\delta}/n) = O(n^{\nu+5\delta/2-1}) = o(n^{-\delta/2}),$$

for sufficiently small δ .

Therefore, the total expected number of offspring of v , in $\mathcal{T}'(w)$, is at most $1 - Kn^{-\delta/2} + o(n^{-\delta/2}) \leq 1 - z$ for $z = (K/2)n^{-\delta/2}$. So $\mathcal{T}'(w)$ grows like a Galton-Watson branching process with branching parameter at most $1 - z$. The probability that such a branching process has size at least x drops quickly as x exceeds $\Theta(z^{-2})$ (see, eg. [6]), and in particular, $\Pr(|\mathcal{T}'(w)| > n^{3\delta/2}/(r-1)) = o(1/n)$. So a.a.s. $|\mathcal{T}'(w)| \leq n^{3\delta/2}/(r-1)$ for every w .

Note that at most $r-1$ Red vertices are formed each time a member of $\mathcal{T}'(w)$ is removed. So a.a.s. the number of Red vertices is at most $(r-1)|\mathcal{T}'(w)| \leq n^{3\delta/2} < n^{2\delta}$ and so $\mathcal{T}(w) = \mathcal{T}'(w)$ for all w . Therefore a.a.s.

$$|T(w) \setminus \mathcal{C}_2| = |\mathcal{T}(w)| \leq n^{3\delta/2}/(r-1) \text{ for every } w \in \cup_{i \geq i_1} S_i.$$

Since each vertex in $T(w) \setminus \mathcal{C}_2$ can be adjacent to at most $r - 1$ vertices in \mathcal{C}_2 , it follows that a.a.s. $|T(w)| \leq n^{3\delta/2}$ for every $w \in \cup_{i \geq i_1} S_i$. This implies that the choice of w below Claim 25 satisfies Lemma 21(b).

Now we prove Claim 25, i.e. we show that a.a.s. each $T(w) \setminus \mathcal{C}_2$ induces a directed tree rooted at w in \mathcal{D} .

Observation: If $T(w) \setminus \mathcal{C}_2$ does not induce a directed tree rooted at w in \mathcal{D} , then there must have been an iteration where some $v \in \mathcal{T}(w)$ is deleted, and one of the $r - 1$ neighbours of v in the remaining hypergraph was either Red or in $\mathcal{T}(w)$.

Again, it will be convenient to consider $\mathcal{T}'(w)$ rather than $\mathcal{T}(w)$.

When we delete v , we choose $r - 1$ uniform vertex-copies as its neighbours. Each vertex of $\mathcal{T}'(w)$ has entered \mathcal{Q} and so has at most one copy remaining. So the probability that we choose a copy of a vertex in $\mathcal{T}'(w)$ is at most $|\mathcal{T}'(w)|/\Theta(n) = O(n^{3\delta/2-1})$. So the probability that this happens during the deletion of at least one member of $\mathcal{T}'(w)$ is $O(n^{3\delta/2}) \times O(n^{3\delta/2-1}) = O(n^{3\delta-1})$.

To bound the probability of choosing a copy of a Red vertex, note that the total number of copies of Red vertices is a.a.s. $O(n^{2\delta} \log n)$, because $\mathcal{T}'(w)$ allows only at most $n^{2\delta}$ vertices to be coloured Red, and a.a.s. the degree of each vertex is less than $\log n$. Since the $r - 1$ vertex-copies (i.e. the neighbours of v) are uniformly chosen from the remaining $\Theta(n)$ vertex-copies, the probability that one of them is a copy of a Red vertex is $O(n^{2\delta-1} \log n)$. Therefore, the probability that during the deletion of vertices in $\mathcal{T}'(w)$, there is such v such that one of its $r - 1$ neighbours was Red is $O(n^{3\delta/2}) \times O(n^{2\delta-1} \log n) = O(n^{5\delta/2-1} \log n)$.

So the probability that, upon deleting some $v \in \mathcal{T}'(w)$, one of the $r - 1$ neighbours is either Red or in $\mathcal{T}'(w)$ is $O(n^{3\delta-1}) + O(n^{5\delta/2-1} \log n) = O(n^{3\delta-1})$. Multiplying by the $O(n^{\nu+\delta/2})$ choices for $w \in \cup_{i \geq i_1} S_i$ (by Lemma 23(c)) we get $O(n^{\nu+7\delta/2-1}) = o(1)$ for δ sufficiently small. This proves that a.a.s. there is no iteration where we delete some $v \in \mathcal{T}'(w)$ and one of the $r - 1$ neighbours of v is either Red or in $\mathcal{T}'(w)$. We proved above that a.a.s. $\mathcal{T}(w) = \mathcal{T}'(w)$ for all w , and so the same is true for $\mathcal{T}(w)$. So our Observation proves that a.a.s. $T(w) \setminus \mathcal{C}_2$ induces a directed tree rooted at w in \mathcal{D} for every $w \in \cup_{i \geq i_1} S_i$, which proves (a,b,c) as described above. □

9 Concluding remarks

We have examined the solution clusters of $X_r(n, cn)$ for $c = c_r^* + o(1)$. We showed that for small constant $\delta > 0$: when $c = c_r^* - n^{-\delta}$, a.a.s. all solutions are $n^{O(\delta)}$ -connected; whereas when $c = c_r^* + n^{-\delta}$, a.a.s. the connectivity parameter of each cluster is $n^{\Theta(\delta)}$, and different clusters are $\Omega(n^{1-r\delta})$ -separated. This indicates a rather smooth cluster transition near the clustering threshold.

It is possible that the clusters are even more separated than we have shown. We would like to know whether the clusters are a.a.s. pairwise $\Omega(n)$ -separated; or all solutions of $X_r(n, cn)$ are a.a.s. $o(n)$ -connected.

When $c = c_r^* + n^{-\delta}$ we have shown that a.a.s. each cluster contains two solutions that are n^{δ} -separated for some constant $z > 0$. We conjecture that this holds as well when $c = c_r^* - n^{-\delta}$ (see Conjecture 8). However, our proof technique for $c = c_r^* + n^{-\delta}$ does not

apply to $c = c_r^* - n^{-\delta}$.

More importantly, we would like to see what happens when $\delta > \frac{1}{2}$ and so we do not know that a 2-core arises *w.h.p.*. In that setting, consider creating a sequence of random hypergraphs by starting with n vertices and adding uniform random hyperedges one at a time. We would like to find out whether clusters arise as soon as the first non-empty 2-core appears in the underlying hypergraph. Is there always some $f(n) = o(g(n))$ such that if there is a 2-core then, under the cluster definitions from Section 2.3, any two solutions in the same cluster are $f(n)$ -connected, while any two solutions in different clusters are $g(n)$ -separated? Or do we need to change the definition of clusters? Or perhaps our notion of clustering falls apart at the very moment when the 2-core is formed.

Instances of r -XORSAT can be solved in polynomial time, using global algorithms such as Gaussian elimination. However, when c is above the clustering threshold, random r -XORSAT seems to be very difficult for generic CSP solvers and local algorithms such as WalkSat[26]. It is natural to wonder whether such difficulties arise precisely at the step when the 2-core appears. Answering this question, and understanding exactly what these difficulties are, could provide insights into how it is that clustering creates algorithmic difficulties for other random CSP's. Resolving the issues discussed in the previous paragraph would be very helpful for this question.

References

- [1] D. Achlioptas and A. Coja-Oghlan. *Algorithmic barriers from phase transitions*. In 49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA, pages 793–802. IEEE Computer Society, 2008.
- [2] D. Achlioptas and M. Molloy, The solution space geometry of random linear equations, *Random Structures & Algorithms* 46(2): 197–231, (2015). arXiv:1107.5550 .
- [3] D. Achlioptas and F. Ricci-Tersenghi, Random formulas have frozen variables. *SIAM J. Comput.*, 39 (1), 260–280 (2009).
- [4] A. Braunstein, M. Mezard and R. Zecchina, Survey propagation: an algorithm for satisfiability, *Random Structures & Algorithms* 27: 201–226, (2005).
- [5] B. Bollobás, The evolution of random graphs, *Transactions of the AMS*, 286: 257–274, (1984).
- [6] B. Bollobás, C. Borgs, J. Chayes, J. Kim and D. Wilson, The scaling window of the 2-SAT transition, *Random Structures & Algorithms* 18: 201–256, (2001).
- [7] Julie Cain and Nicholas Wormald, Encores on cores, *Electron. J. Combin.*, 13(1), Research Paper 81, 13 pp, (2006).
- [8] N. Calkin, Dependent sets of constant weight binary vectors, *Combinatorics, Probability and Computing*, 6: 263–271, (1997).

- [9] V. Chvátal, Almost all graphs with $1.44n$ edges are 3-colorable, *Random Structures & Algorithms*, 2(1): 11–28, (1991).
- [10] S. Cocco, O. Dubois, J. Mandler and R. Monasson, Rigorous decimation-based construction of ground pure states for spin glass models on random lattices. *Phys. Rev. Lett.* 90, 047205, (2003).
- [11] A. Coja-Oghlan, A better algorithm for random k -SAT, *SIAM Journal on Computing* 39: 2823–2864, (2010).
- [12] A. Coja-Oghlan and C. Efthymiou, On independent sets in random graphs, *Proc. 22nd SODA*, 136–144, (2011).
- [13] A. Coja-Oghlan, K. Panagiotou, Catching the k -NAESAT threshold, *Proc. 44th STOC*, 899–908, (2012).
- [14] A. Coja-Oghlan, K. Panagiotou, Going after the k -SAT threshold, *Proc. 45th STOC*, 705–714, (2013).
- [15] A. Coja-Oghlan, D. Vilenchik, Chasing the k -colorability threshold, *Proc. FOCS*, 380–389, (2013).
- [16] A. Coja-Oghlan and L. Zdeborova, The condensation transition in random hypergraph 2-coloring, *Proc. 23rd SODA*, 241–250, (2012).
- [17] A. Coja-Oghlan, The asymptotic k -sat threshold, *Proc. STOC*, 804–813, ACM, (2014).
- [18] N. Creignou and H. Daudé, Satisfiability threshold for random XOR-CNF formulas, *Discrete Appl. Math.*, 96-97: 41–53, (1999).
- [19] H. Daudé, M. Mézard, T. Mora, and R. Zecchina, Pairs of SAT assignments and clustering in random boolean formulae, *Theoretical Computer Science*, 393(1-3): 260–279, (2008).
- [20] M. Dietzfelbinger, A. Goerdt, M. Mitzenmacher, A. Montanari, R. Pagh and M. Rink, Tight thresholds for cuckoo hashing via XORSAT, *Proceedings of Automata, Languages and Programming, 37th International Colloquium, ICALP*, 213–225, 2010.
- [21] O. Dubois and J. Mandler, The 3-XORSAT threshold, *Comptes Rendus Mathématique*, 335(11): 963–966, (2002).
- [22] D. Gamarnik and M. Sudan, Limits of local algorithms over sparse random graphs, *Proceedings of the 5th conference on Innovations in theoretical computer science* 369–376, ACM, (2014).
- [23] J. Ding, A. Sly, and N. Sun, Proof of the satisfiability conjecture for large k , arXiv:1411.0650, (2014).
- [24] P. Gao and M. Molloy, *The stripping process can be slow: part I*, arXiv: 1501.02695.

- [25] P. Gao, *The stripping process can be slow: part II*, arXiv: 1505.02804.
- [26] M. Guidetti and A.P. Young, Complexity of several constraint-satisfaction problems using the heuristic classical algorithm WalkSAT, *Phys. Rev. E*, 84 (1), 011102, 2011.
- [27] M. Ibrahimi, Y. Kanoria, M. Kramlinger, and A. Montanari, The set of solutions of random xorsat formulae, *Proc. SODA*, 760–779, SIAM, 2012.
- [28] S. Janson, T. Łuczak, D. Knuth, and B. Pittel, The birth of the giant component, *Random Structures & Algorithms* 3: 233–358, (1993).
- [29] J.H.Kim, Poisson cloning model for random graphs, *International Congress of Mathematicians* Vol. III, 873–897, Eur. Math. Soc., Zürich, 2006.
- [30] V. Kolchin, Random graphs and systems of linear equations in finite fields, *Random Structures & Algorithms*, 5: 135–146, (1994).
- [31] V. Kolchin and V. Khokhlov, A threshold effect for systems of random equations of a special form, *Discrete Mathematics and Applications*, 5: 425–436, (1995).
- [32] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborova, Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems, *Proc. Natl. Acad. Sci.*, 104(25): 10318-10323, (2007).
- [33] T. Łuczak, Component behaviour near the critical point of the random graph process, *Random Structure & Algorithms*, 1: 287–310, (1990).
- [34] T. Łuczak, B. Pittel and J. Weirman, The structure of a random graph at the point of the phase transition, *Trans. Am. Math. Soc.* 341: 721–748, (1994).
- [35] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., New York, NY, USA, 2009.
- [36] M. Mézard, T. Mora, and R. Zecchina, Clustering of solutions in the random satisfiability problem, *Phys. Rev. Lett.*, 94(19), 197205 (2005).
- [37] M. Mézard, G. Parisi, and R. Zecchina, Analytic and algorithmic solution of random satisfiability problems, *Science*, 297: 812–815, (2002).
- [38] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina, Two solutions to diluted p -spin models and XORSAT problems, *J. Stat. Phys.* 111: 505–533, (2003).
- [39] M. Mezard, R. Zecchina, The random K -satisfiability problem: from an analytic solution to an efficient algorithm, *Phys. Rev. E* 66, (2002).
- [40] M. Molloy, Cores in random hypergraphs and boolean formulas, *Random Structures & Algorithms*, 27: 124–135, (2005).
- [41] M. Molloy, The freezing threshold for k -colourings of a random graph, *Proc. STOC*, 921–930, (2012).

- [42] M. Molloy and R. Restrepo, Frozen variables in random boolean constraint satisfaction problems, *Proc. SODA*, 1306–1318, (2013).
- [43] A. Montanari, R. Restrepo and P. Tetali, Reconstruction and clustering in random constraint satisfaction problems, *SIAM J. Disc. Math.*, 25: 771–808, (2011).
- [44] R. Mulet, A. Pagani, M. Weigt and R. Zecchina, Coloring random graphs, *Phys. Rev. Lett.* 89, 268701, (2002).
- [45] B. Pittel and G. Sorkin, The satisfiability threshold for k -XORSAT, arXiv:1212.1905 (2012).
- [46] B. Pittel, J. Spencer and N. Wormald, Sudden emergence of a giant k -core in a random graph, *J. Comb. Th. B* 67: 111–151, (1996).
- [47] G. Semerjian, *On the freezing of variables in random constraint satisfaction problems*. J.Stat.Physics **130** (2008), 251 - 293.
- [48] L. Zdeborová and F. Krzakala, Phase transitions in the colouring of random graphs, *Phys. Rev. E* 76, 031131, (2007).